Re-examining the SLIP Task: Completely Lexical vs. Completely Non-Lexical

Shannon East

MSc Psycholinguistics
University of Edinburgh
2007

Abstract

For years, the SLIP task, a methodology used to elicit speech errors in a laboratory setting, has added invaluable evidence in developing models of language production. It has lead to the finding that speech errors more commonly result in real words than in non-words, called the 'lexical bias effect'. In turn, it has been used to support the idea of a self-monitor that monitors for speech errors, and is more likely to let real words slip past than non-words, as well as a theory of feedback between levels of processing, which increases the number of lexical errors, and more recently a theory that combines the self-monitor and feedback. However, past usage of the SLIP task has never directly compared results for a completely lexical context to that of a completely non-lexical context. The purpose of the present study is to fill this gap in the literature by using two levels of context, one block where no non-words are presented, and another where no real words are presented, as well as an outcome condition, where the intended error outcome of a target is either lexical or non-lexical. 52 participants read word pairs and the number and variety of speech errors and other responses they made were recorded. The results found no significant evidence for a difference in the number of exchanges between any conditions, however there were effects for the number of corrections made and for the number of times participants failed to respond. These results can add support to existing models of language production, particularly in the role of feedback in the selection of lexical concepts after an error is detected by the monitor.

Table of Contents

1. Introduction

1.1 Lexical Bias and Errors

The "lexical bias effect" is the predisposition for phonological slips to result in real words more often than in not real words (Baars, Motley, & Mackay, 1975; Dell, 1986, 1990; Dell & Reich, 1981; Hamm, Junglas, & Bredenkamp, 2004; Humphreys, 2002; Nooteboom, 2005). For example, if you intended to say 'a pack of lies' the real-word outcome 'a lack of pies' would be more common than producing an error that is just a non-lexical string of sounds. This effect has been shown in real-world observations, corpus studies, as well as laboratory experiments, although not all evidence has been in support of the effect (see Del Viso, Igoa, & García-Albea, 1991; Garrett, 1976).

1.2 Models of Production

Levelt, Roelofs and Meyers (1999) developed a model of speech production that relies on a monitor. In their model, there are multiple levels that feed forward to produce speech (see Fig. 1). It starts with conceptual preparation and from that a lexical concept, so if you are trying to produce the word 'sponge' you will initially only have a concept of spongy-ness: you might have a vague image of the thing sitting on your kitchen sink or of the porous marine filter-feeders, devoid of any other features such as morphology or phonology. Then the lemma, a more concrete idea of the word with more information for later encoding than the lexical concept, is retrieved from the mental lexicon, and you have a more concrete idea of what your sponge is, and are ready to turn that idea into a word. From there the process goes through different levels of encoding to produce the appropriate morphemes. In this case, it is a mono-morphemic word so just a representation of the morpheme 'sponge', phonology, /s p Λ n ĵ/, and syllables, on to the gestural score for how to articulate the word, with information on how to move your mouth and such parts to produce the desired phonemes, and finally the word is produced by the vocal tract, and you can declare your thoughts on some such a member of the phylum Porifera. In this model, spoonerisms and other errors would be recognized by a monitor. There is an allowance for an internal as well as an external monitor, the external would monitor overt speech, and the internal monitor that can detect errors as early as

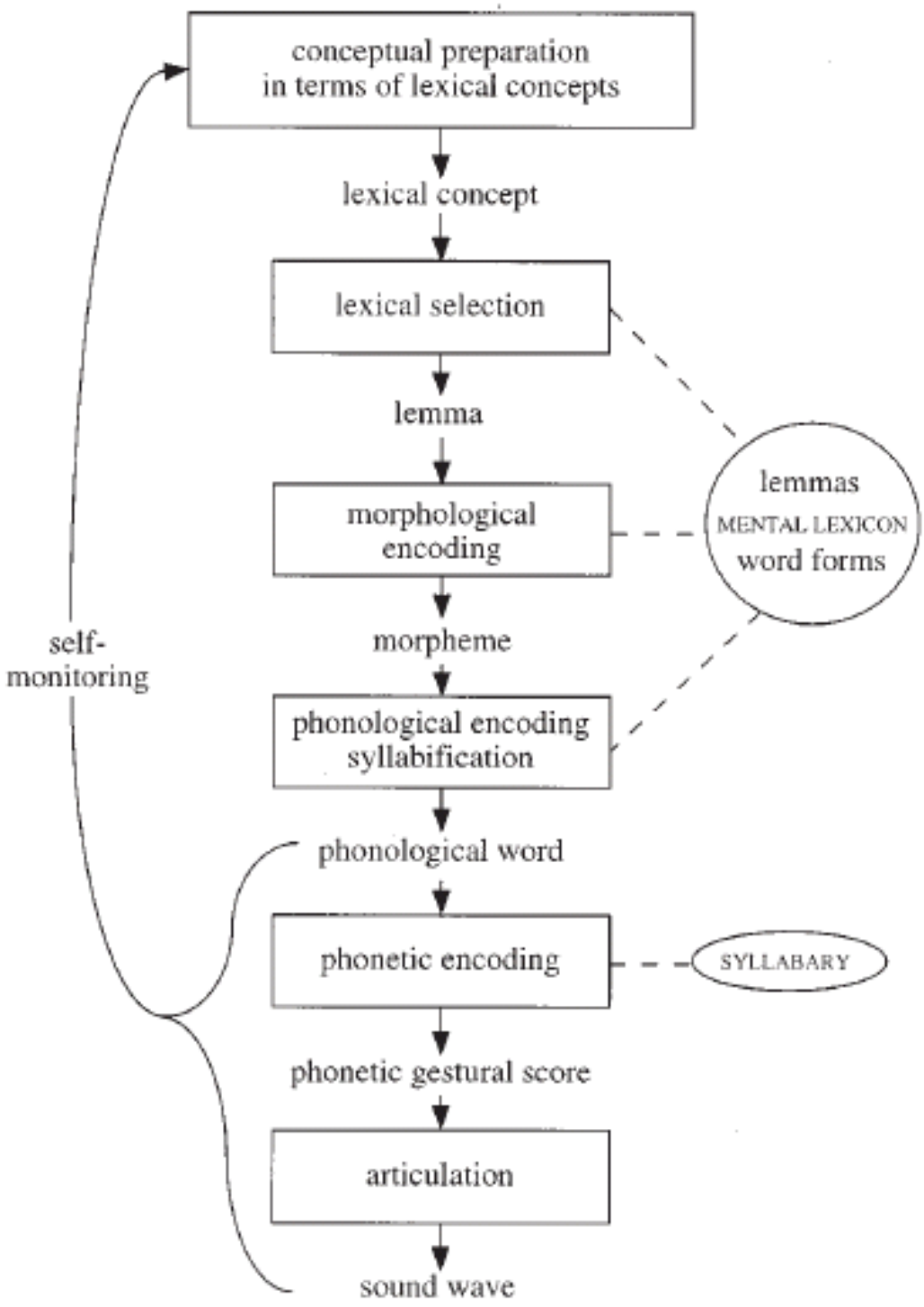

Figure 1: The Theory in Outline: Preparing a word proceeds through stages of conceptual preparation, lexical selection, morphological and phonological encoding, and phonetic encoding before articulation can be initiated. (Levelt et al., 1999, p. 3)

phonological encoding, before phonetic encoding, and continues to look for errors until the word is articulated and it switches the task over to the external monitor. If the monitor does find an error, then the procedure has to star again back at the conceptual preparation level.

Dell also added to the debate of how speech is produced with information from the study of speech errors through his spreading activation model of language production (1986). His claim was that spreading activation is not just limited to occurring within a layer of processing, but that it can result in feedback between the different levels of speech and can propagate back up to previous levels of activation. This feedback can be used to repair errors that are detected more efficiently than with the monitor, because it just requires feedback up to the previous level instead of having to start the production process all over again, as you would with the monitor. Feedback also is useful in explaining phonological priming, because when spreading activation activates the phonological units, it would then propagate back up to lexemes that share phonetic features. So, in the context of the lexical bias effect, this would mean that errors at the phonological level that result in real words are less likely to be corrected because the feedback to the morpheme level would only confirm that a lexical error is indeed a real word. For example, if you were trying to say the word 'kitten' and instead the activation was higher for the phoneme /*m*/ than /*k*/, possibly due to increased activation due to one of the surrounding words, the feedback up to the morphemic level would just confirm that 'mitten' is a real word, and the activation back down would remain stronger for 'mitten'. Conversely, if /*d*/ had a higher activation level than /*k*/, the feedback would not be able to activate another plausible morpheme as a real word, and the error would not be as likely to propagate itself (example taken from Dell, 1986).

1.3 Testing the Lexical Bias

One particular method of eliciting speech errors that served as the earliest demonstration of the lexical bias in an experimental setting is the Spoonerisms of a Laboratory Induced Predisposition, or SLIP, procedure. First developed by Baars and Motley in 1974 and most notably used by Baars, Motley and MacKay (1975), it was designed to test whether the lexicality of the context and the lexicality of the potential errorful outcome impacted the error rate. It utilized priming to try to lead participants

to being more error-prone. The stimuli consisted of words or non-words presented in pairs. Within this presentation, some of the words are biasing pairs that are intended to prime the targets so participants would make a speech error on the target pair by using the same onset as the intended error outcome. For example, the participants would see word pairs along the lines of 'lap fail' proceeding the target 'fate lame', in the hopes that the target would spoonerize into 'late fame'. Participants were told that they would hear a buzzer, which was a cue to say out loud the word pair they had just seen. In this experiment Baars et al. manipulated the lexical context, so that the word pairs were presented in blocks that either consisted entirely of non-words, or in blocks with a 'mixed context' where half of the filler word pairs in the block were real words, to provide a more lexical context (Baars et al. refers to this context as 'lexical' but this paper will refer to it as 'mixed' to avoid confusion with a context comprised of entirely real words). They found that in the mixed context, there was a large disparity between the real-word outcomes and the non-word outcomes, with a much higher rate of errors for real-word pairs, which resulted in the lexical bias effect, although in the non-lexical context, this effect was not present.

The lexical bias effect that varied by lexical context seen in Baars et al. (1975) could be counted as evidence for Levelt's monitor, but cannot be so clearly explained by Dell's feedback. In the model of feedback, there is no particular reason why feedback should impact the number of errors by context. However, a smart monitor could be able to discern the lexical context of the intended utterance, and essentially stop monitoring if it knows that none of the utterances could be lexical.

More recently, Hartsuiker, Corley and Martensen (2005) suggested that a model of speech production should incorporate both feedback and a self monitor. They too used the SLIP task, although this time ran a more tightly controlled version of Baars et al. (1975), making sure to control for spoonerisms that were not just phonetically real words, but also orthographically real words. They employed a mixed and a non-lexical context, both of which had non-lexical targets that could spoonerize into either real or not real words, and in the mixed context half of the fillers were real words, all the other word pairs were non-words. They found that the lexical bias did persist in that targets with lexical outcomes, when spoonerized, were more common than those that resulted in non-words. This only seemed to be the case in the mixed context, not the non-lexical context, and they also saw a lower number of errors in the non-lexical context than in the mixed context

condition. The authors claimed that the pattern of results was due to both feedback and a smart monitor that can adjust to context. The smart monitor can discern the lexicality of the context and from there decide whether non-words or real words should be considered errors. In the non-lexical context, the smart monitor would count real words as errors because nothing should be a real word. However, in the mixed-lexical context, the smart monitor cannot decide what the output should be. Enhancing the effect is that they claim feedback would increase the number of lexical errors made, because the phonological errors that produce real words will be reinforced through spreading activation, when the error produces a real word it is confirmed that there is a morpheme that matches, and so the error is allowed to propagate back down, but it would not benefit from that increased activation if the error produced a non-word because there would be no matching morpheme to activate. So, in summary, feedback increases the number of lexical errors made, as per Dell's (1986) model, while the monitor would decrease the number of non-lexical errors made.

One thing that has not been fully addressed in the literature is the use of a fully lexical context and a fully non-lexical context. Some experiments use a mixed lexical context and compare it with a non-lexical context (i.e. Baars et al., 1975; Hartsuiker et al., 2005), or use all lexical stimuli throughout the experiment (i.e., Dell, 1984, 1986; Nooteboom, in press), but there has yet to be a paper in the published SLIP literature that has used one block where the participant sees only real words and one block where none of the words they see are lexical. If we take Hartsuiker et al.'s (2005) model of a smart monitor that can adjust to context, it is only fair to look at whether it is simply the mere presence of real words that is enough for the monitor to treat everything as potentially lexical, or if the degree of lexicality in the context is important.

This investigation intends to fill this gap by examining how a fully lexical context is treated. If the smart monitor operates as outlined in Hartsuiker et al. (2005), then the prediction is that the all-lexical context would perform largely the same as the mixed-lexical context. The smart monitor should realize that it is a fully lexical context, and therefore that it should not let the system produce a non-word. The lexical bias should be more pronounced, because the monitor would suppress even more of the non-word error outcomes, and the real-word error outcome rate should be higher, due to more real words being activated through feedback.

1.4 Problems with the SLIP task

The SLIP task, however, does have problems that need to be resolved. Foremost is that, to be effective, the SLIP task needs to elicit a high number of errors. If the error rates are low, then to have any statistical power you would have to run an unreasonable number of participants. Unfortunately, there has been an observable decrease in the error rates since the original runs of the procedure. Baars et al. (1975, Experiment 2) recorded that 8.2% of the target responses were full exchanges, but the highest since then has only been 5% (Humphreys, 2000, Experiment 3), and the lowest error rate was 0.45% (Dell, 1990, Experiment 2).

So what can be done about the power problem, and what have researchers changed that would cause this dramatic drop-off? The most significant change has come with the transition from mechanical stimuli to computerized stimuli. When Baars et al. (1975) originally ran their experiment they used a device called a memory drum. This apparatus was, as the name suggests, a large cylindrical drum with a small window for the stimuli to be presented through, which has been around since 1887 when it was first used for verbal memory and learning tasks (Haupt, 2001). The stimuli were on a rotating mechanism inside the drum that would move at a given interval, thus changing the stimuli that appeared. But with technological advancements, computers became a far more practical way of displaying the stimuli, so after the 1980's, experimenters no longer had need of a memory drum. It would appear that with the change of apparatus came a decrease in error rates. But how could the switch to computers make people less prone to errors?

Fundamentally, the computers do the same thing as the memory drum: They present the words in pairs at given time intervals, and the participant is cued to repeat certain pairs. So the change is not in the task, it has something to do with the operation of the device itself. One possibility is the fact that the memory drum was a much noisier device than the computers, and with the change in display came a rather large noise. As described by Baars:

> "Informal observation suggests that the rather loud, regular relay click of the memory drum may serve to pace the subject's speech, thereby increasing the slip rate. This can be simulated on a microcomputer by a brief 0.1-second

> tone, sounded simultaneously with each change in display." (Baars, 1992, p. 133)

One possibility for why this would have an effect comes from Allen (1972, 1975), and his research into speech rhythms. It was found that when people align their speech to a rhythm, they are prone to aligning what has been called the "perceptual center," or "P-center," to the beat. While other research has shown that an individual's perception of where the beat lies within the rhythm can vary from person to person, Allen (1972) found that the P-center remained constant between individuals: at the onset of the nuclear vowel. While there can be variability from word to word, depending on the number of consonants that proceed or follow the vowel, the SLIP task uses simple CVC(C) pattern words, words with a consonant in the initial position, a vowel in the nucleus, and another consonant or two in the coda, which would keep a fairly constant P-center location. At most, the extra consonant at the end of the word would shift the P-center slightly more to the center of the vowel. So, when taken in the context of the memory drum's loud, regular ticking sound, this would provide a rhythm that the participants are more inclined to align their speech with. And in aligning their speech to the rhythm, it draws their focus to the nucleus of the word, and thusly away from the onset, where the spoonerism would come into play.

In summary, this research will primarily compare a fully-lexical context in the SLIP task with a fully non-lexical context, and secondly will attempt to boost the overall error rate with a ticking sound timed to the presentation of the stimuli. This will be done in a close replication of Hartsuiker et al.'s (2005) study. It is predicted, because the monitor is most affected by context, and the monitor suppresses non-lexical error outcomes in the mixed context in past experiments, that the number of non-lexical error outcomes in the lexical context will be lowered if the monitor is sensitive to the amount of lexicality. It is also hoped that the ticking will raise the error rate to being closer to that obtained in Baars et al. (1975).

2. Method – Pilot

2.1 Participants

10 participants were recruited from the area surrounding the University of Edinburgh. They ranged in age from 18 to 34, with a mean age of 25. 7 were female, 3 were male, and all had at least entered some university education.

2.2 Materials

Two lists of 500 word pairs were constructed, one list consisted only of non-word pronounceable letter strings, the other consisted of real words, and all were of the form CVC(C). Each list consisted of 24 target pairs, 120 biasing pairs, 26 control pairs and 330 fillers. Target pairs were the items the participants were to say aloud, in the hopes that they would produce a spoonerism. The spoonerism would result because the target had been preceded by biasing pairs, which were pairs of words or non-words that shared the same initial consonant as the target pair, but with the initial consonants switched within the pair, to prime the spoonerism in the target. For example, the target pair 'kin bit' would be preceded by the biasing pairs bane keen, ball keel, both keep, big kill and bib kick. In both lists, the nuclear vowel would be shared within the target pair, as well as with the two biasing pairs immediately preceding the target. The control pairs were word pairs that were signaled for the participant to say, but did not have any biasing pairs before it. Filler pairs were presented to further establish the context, as well as make it more difficult to predict which pairs were cued for response.

Each block could be broken down into 25 sequences, including a practice sequence, consisting of 20 word pairs: 13 randomized fillers, 1 control pair, 5 biasing pairs, 1 target. Other than keeping the biasing pairs within 7 pairs in front of the target, the order within each sequence was randomized. While the fillers were random across all sequences, it was ensured that within a sequence, none of the fillers shared an initial consonant with the target. The practice sequence, which was the first sequence seen in a block, either when starting the experiment or when starting the second half, after a break, had 3 controls and 17 bias pairs, presented randomly.

2.3 Procedure

Each participant was evaluated individually in a quiet room. They wore headphones throughout the experiment, and through the headphones they heard white noise, loud enough to block out the surrounding sound but not so loud that it was uncomfortable. Participants were given instructions to read the word pairs silently as they were presented, and upon hearing a beep they were to say the previous word pair out loud as quickly as possible, with their response recorded using a microphone. The sound of their voice would trigger moving on to the next slide, but if they did not respond quickly enough a red screen would appear as a warning. After reading the instructions, the experiment began, starting with the practice sequence. The lexical and non-lexical contexts were presented in blocks, and which block came first was counterbalanced across subjects.

Each word pair was presented for 700 ms, followed by a blank screen for 200 ms before moving to the next pair. When the target or control pairs were presented, a beep would be heard immediately after the screen was cleared, which was the signal to say the last pair aloud. If the participant failed to begin to say the word after 500 ms, a red screen appeared and they heard a louder sound, distinct from the signal beep, as an indicator that future responses should be made faster.

Responses were recorded directly into a sound file, with a sample rate of 96 kHz and analyzed offline. As in Baars et al.'s (1975) and Hartsuiker et al.'s (2005) studies, each utterance was coded as either a correct utterance (*beck weld*), a full exchange (*weck beld*), a partial exchange, including either anticipations (*weck weld*) or preservations (*beck beld*), an other error, or as a failure to respond. However, in line with Nooteboom and Quené (in press), an additional category was added for interrupted exchanges, where the participant initiates an error but either aborts or corrects themselves prior to completion (*weck* or even *we*, prior to stopping or restarting correctly), as well as a category for competing errors, where the initial consonants are exchanged, but the rest of the word pair does not match the stimuli (*wealth best*).

3. Method- Main Study

3.1 Participants

42 participants were recruited from the area surrounding the University of Edinburgh, none of whom had taken part in the pilot. They ranged in age from 19 to 50, with a mean age of 23. 27 were female, 15 were male, and all had at least entered some university education.

3.2 Materials

The materials were the same of the pilot, as described in Section 2.2, with one exception: in an attempt to boost errors rates, a ticking sound was played with the presentation of each word pair. This sound was a cow bell sound produced by a Roland XP-10 Synthesizer, 120 ms in length, and was heard whenever a word pair was presented, which meant the onset would be heard 900 ms after the onset of the last sound, except between the target pair and the following word pair, which could vary due to the transition's sensitivity to the participant's voice onset when speaking the target pair.

3.3 Procedure

The procedures for the main study were identical to that of the pilot experiment, as described in Section 2.3.

4. Results

When examining the pilot's ten participants, they made 461 total responses to target items, including 31 exchanges (6.72%), of which 18 were full exchanges and 13 partial exchanges, also 17 interruptions, and 41 other errors. Looking at the data from the 42 participants who heard the ticking that was not present in the pilot, there were 1975 target responses, of which 100 were exchanges (5.06%), including 74 full exchanges, 26 partial exchanges, as well as 33 interruptions, and 17 other errors. Out of 2436 total target responses, there were 131 exchanges, or 5.38% of target responses, of which 39 were partial exchanges, and 92 were full exchanges, as well as 46 interruptions and 58 other errors (see Table 1). No competing errors were made on any of the target stimuli.

Table 1: Distribution of errors.

	Lexical outcome				Non-lexical outcome			
Context	Exchanges	Full	Partial	Corrections	Exchanges	Full	Partial	Corrections
Lexical	36	23	13	18	33	23	10	16
Non-word	31	26	5	8	31	20	11	4

An analysis of variance (ANOVA) was performed to examine the significance of the different interactions for the different categories of errors. They were performed with Context (either lexical or non-lexical) and Outcome (lexical or non-lexical) as within-subjects factors. When looking at the total exchanges, which includes both full and partial exchanges together, Context was not significant, $F<1$, and neither was Outcome, $F<1$, or even Context by Outcome, $F<1$, (see Fig 2).

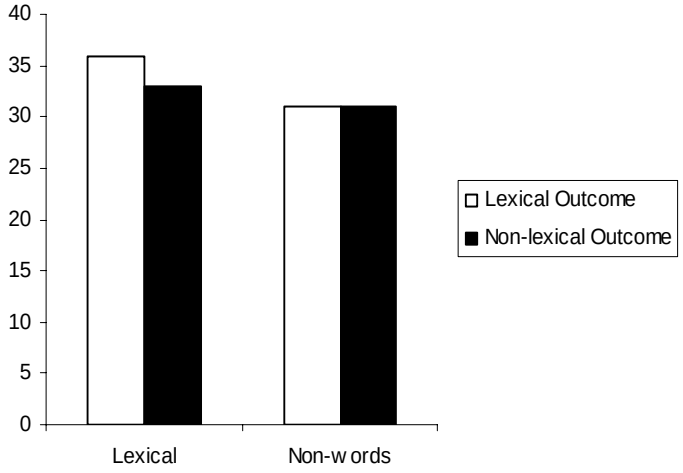

Figure 2: Number of Full and Partial Exchanges

When examined separately, the full exchanges alone also saw no significant effect of Context, $F<1$, Outcome, $F<1$, or Context by Outcome, $F<1$. Partial exchanges saw no effect of Context, $F<1$, or Outcome, $F<1$, and a marginally significant interaction of Context by Outcome, $F(1, 51) = 3.705$, $p > .05$, $\omega^2 = .060$.

For interruptions there was a significant effect of Context, $F(1, 51) = 11.119$, $p < .05$, $\omega^2 = .002$, but not Outcome, $F<1$, or Context by Outcome, $F<1$.

A rather unexpected turn came with the analysis of when participants did not respond. Context alone showed no significance, $F(1, 51) = 2.014$, $p > .05$, $\omega^2 = .162$, and neither did Outcome alone, F<1. However, there was a significant Context by Outcome interaction, $F(1, 51) = 8.088$, $p < .05$, $\omega^2 = .006$, and would appear that

when the lexicality of the context matches that of the outcome, the speaker was more likely to not respond (see Fig 3).

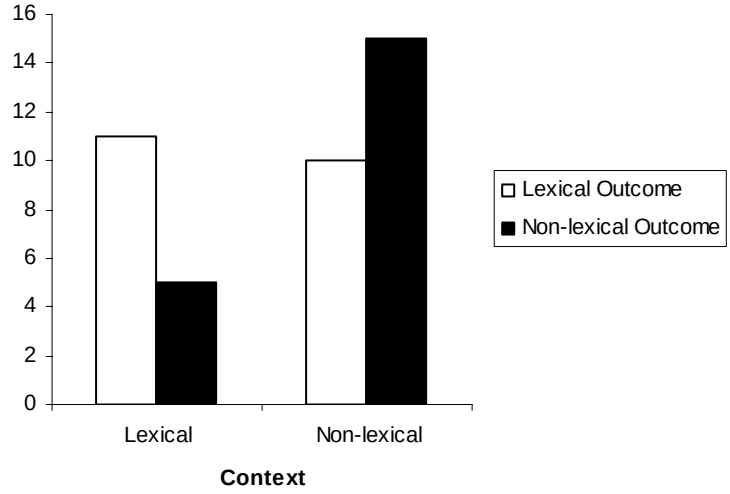

Figure 3 Number of Times Participants Failed to Respond to a Target

5. Discussion

While the prediction that there would be a stronger effect of lexical bias in a fully lexical condition was not verified, three primary issues are raised by the results of this experiment. Foremost is the lack of significant results for exchanges. The significance of context in the number of interruptions also raises some questions, as does the lack of competing errors. While the significance of context by outcome for the non-response category is not a result that has been reported before, this and the other results can be explained in terms of existing production models.

5.1 Null Results

There are a number of possible explanations for the null results, even without experimental design flaws. The experimental procedure was a nearly exact replication of Hartsuiker et al. (2005), and while there were changes to the lexicality of one context and to the background noise during the main experiment, the results for the non-lexical context would be expected to be the same, so the null result is a bit surprising. The main difference in methods was that this experiment used more people, each of whom produced more errors on average. There was a good deal of variability from person to person: some people made no errors, some people made

errors very regularly, including in response to the un-primed control items. Overall, it appeared that the additional people making more errors, while it would seem beneficial, only leveled out the pattern of results, in spite of or perhaps because of the increase in statistical power.

The 'file drawer' phenomenon could also contribute to why this experiment's findings do not appear to fit in with the existing literature. Experiments that cannot disprove the null hypothesis are less likely to get published, and instead are put away in the back of the file drawer, never to be heard from again. The upshot of this, however, is that the literature cannot accurately show if a given method is really valid, because if 95 times out of a hundred a methodology produces null results, and only those five that saw results were published, the literature will skew towards the significant results, even if they were just a lucky fluke. The 95 apparently failed attempts could have been due to procedural problems that need to be addressed, or they equally could have been attempts at fixing procedural problems from the experiments that found results, only to find that the results are negated, but the published literature will not reflect any of this. There are roughly XX successful publications of experiments using the SLIP procedure. It seems unlikely that, in the 32 years since its famous first use, that the relatively small number of runs we see in the published literature represent the full number of the times the SLIP task has been used, or even a majority of the replications. And even within those runs, the slipping error rate brings to question how generalizable the results may be.

5.2 Interrupted and Competing Errors

The significance of context for the interruptions is surprising in light of Nooteboom and Quené's (in press) results, although the complete lack of competing errors is not. Their experiment did not use context as a variable, and they aimed to look at where in time different types of errors occurred. They found that interruptions are a result of earlier phases of processing, and the competing errors come from later phases. The lack of competing errors is most likely due to the shorter response deadline: they found that the response time was generally later for competing errors, so the time pressure in this experiment prevented participants from making competing errors. The contextual difference for interruptions could be the result of the monitor being more active in the lexical context, so that the auditory loop can more readily recognize that the production is inconsistent with what it ought to be. It is also

possible, because Nooteboom views interruptions as 'hasty' errors, that the certainty of the lexical context, that is knowing that anything you say should be a real word, results in people articulating sooner, before the internal monitor had a chance to detect the error, whereas in the non-lexical context, because nothing is as certain with a non-lexical context and the inactive monitor, the process is slowed down a bit more. This possibility, however, could not be verified in this experiment, as the reaction times were not recorded.

5.3 Non-Response

The issue presented by the 'no response' category is a tricky one. The primary question revolves around what the lack of response could be interpreted to mean, what processing went on behind it to cause such an error. Clearly it is, in some form, a failure of the language production system, as no language is produced. But the lack of output makes it all the more difficult to discern where in any language production model the failure takes place at, as such models seem appropriately concerned with the production of language, not the lack of production. It seems most likely that the problem occurs at the conceptual level, either during initial selection or after an error is detected and selection at the conceptual level has to be made again.

There are a couple different plausible explanations for why there would be production difficulties during the initial conceptual selection. Memory would seem to be a likely explanation for why it would have difficulty starting in the first place: the participant just forgot what the concept or phonemic cluster they were supposed to articulate was in the first place. This, however, would not explain the significant difference for context by outcome. It could also be a problem of selection difficulty, where the speaker has difficulty choosing the lexical concept they intend to articulate, especially when under a time pressure, and can result in a disfluency, particularly errors or filled pauses like 'um' and 'uh' (Oomen and Postma, 2001).

From here it could be that the context by output difference could occur because of instances when the monitor has sent the process back to the beginning. If some form of error is detected once encoding has begun, the existing models (i.e., Dell, 1986; Levelt and Roelofs, 1999; Postma, 2000; Hartsuiker et al., 2005) indicate that either it is detected by a monitor, and processing returns to the conceptual level, or the error is undetected and goes on to be produced. Postma (2000) outlines eleven distinct feedback loops that can come into play (see Figure 4), and puts them into

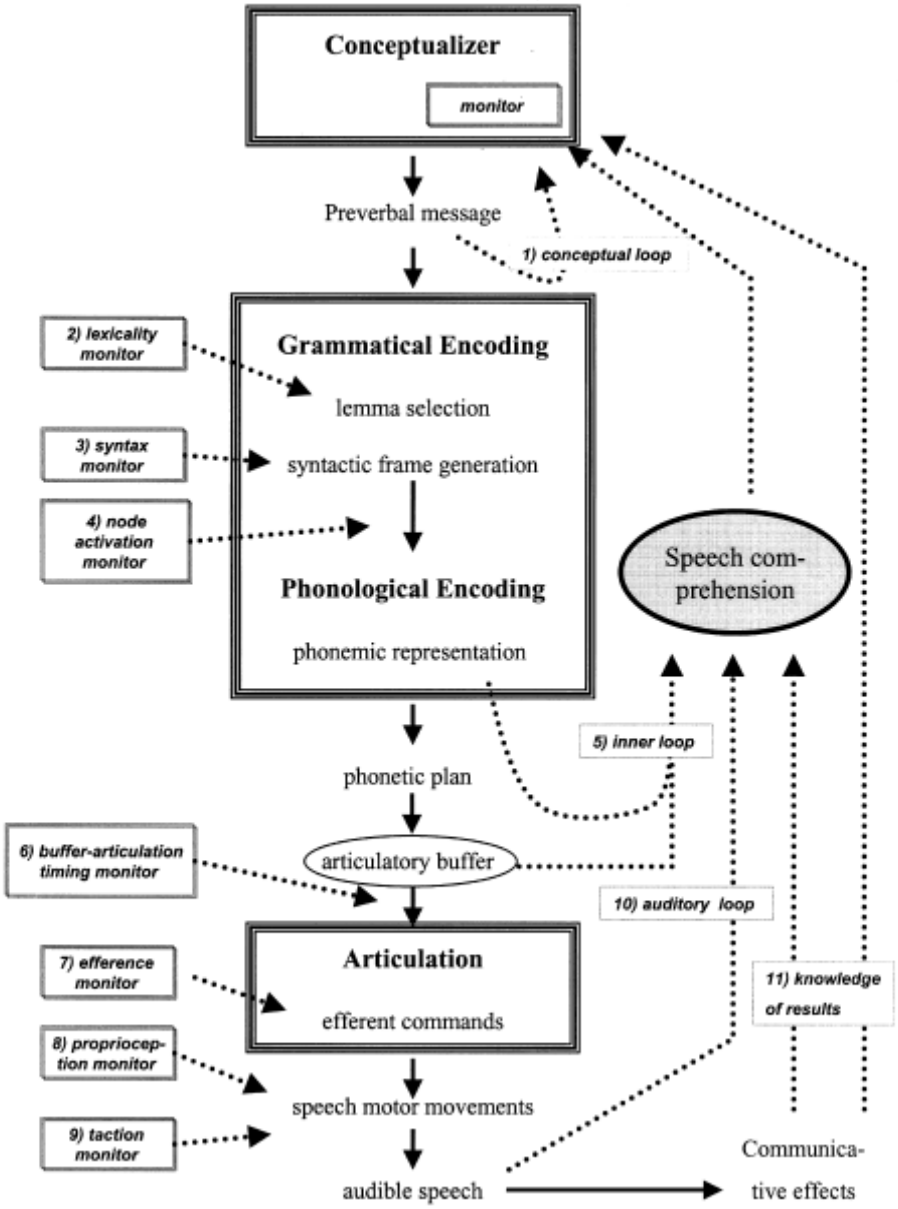

Figure 4: Model of speech production, with a break down of the different loops that comprise the monitor (Postma, 2000, p. 99)

three broader categories: intrinsic feedback, response feedback, and external feedback. The first category, intrinsic feedback, seems to be where the error would have to occurred in the case of a non-response: as nothing is produced, it cannot be a problem of the external monitor, and the response feedback loops involve feedback within the peripheral nervous system, so the response feedback loop is unlikely to be at all effected by the context and outcome conditions. That leaves the seven initial loops in the intrinsic feedback category. If it is not a problem of encoding or articulation, then that leaves only the pre-verbal conceptual levels, or possibly the

monitor itself, as the guilty party. Blackmer and Mitton (1991) and Van Hest (1996) report that conceptual repairs take place more slowly than repairs to other forms of errors at lower levels. It is postulated that this is because it is either harder to select the proper repair or to reject the incorrect response at this level, although it appears to be impossible to experimentally differentiate between the two possible explanations.

Why, then, would difficulty with the selection of a lexical concept after recovery from an internal speech error show a context by outcome effect? One possibility is that it becomes a problem of options: that selection difficulty becomes a bigger problem when it is trying to recover from an internal error, because spreading activation after an error will impact the options there are to select from. When production goes through the different stages of processing, Dell (1986) models that it is a competition between the activation levels of different nodes. Because of spreading activation, within each level, not only is the intended outcome activated, but all the nodes that are connected to this node also become at least slightly more activated as a result. This is how a speech error could be made: a preservation, for example saying 'pipe poke' instead of 'pipe smoke', would result from when the node for /*p*/ at the phonological level still has a higher-than-baseline level of activation from the previous pronunciations, and then spreading activation adds to the total activation, making it more highly activated than the node for the real intended outcome, in his case /*s*/ and /*m*/. But then once this node is selected, the activation would propagate back up through the other levels of activation, so then the morpheme, lemma, and lexical concept for 'poke' become more highly activated than other nodes on each given level.

The results from the present study show that you are more likely to omit an answer when the lexicality of the outcome matches that of the context. In the lexical context with a lexical outcome, if there was an error to begin with that would have caused the speaker to produce another lexical word pair, the feedback could have propagated up to the topmost levels, creating too many options of what to say, and the speaker essentially would be overwhelmed or would give up. In the case of the non-lexical context with a non-lexical outcome, no real words are produced, so when the monitor goes back to the top, there are no lexical concepts activated, creating the selection difficulty at the conceptual level because the production system could not have any concrete form of the word until the phonemic level, and there are no options that were more highly activated because there were no lexical concepts to begin with,

so any activation at higher levels is just due to spreading activation. The lack of lexical concepts from the start could also explain why there was a nearly significant effect by context: with a dearth of lexical options activated on the conceptual level, overall there was just more difficulty selecting a non-lexical production. The difference between the two levels of outcomes was nearly the same for each context, so it is not inconceivable that, while the context did not show a fully significant effect, it does contribute to the overall pattern of results. In both the lexical context/non-lexical outcome as well as the non-lexical context/lexical outcome, it is much easier for the production process to move on, because when it does have to re-select a lexical concept, there is only one highly activated lexical outcome, because the other likely outcome has no lexical concept. Overall, however, as this is the first time that the results for the 'no response' category have been reported, further exploration into this phenomenon is necessary.

6. Conclusions

This investigation into how the lexical bias would be impacted by a fully lexical context in the SLIP task, despite being a close replication of Hartsuiker et al. (2005), did not show a significant effect of condition by outcome, however it still achieved interesting results. While unexpected, the pattern of results found in this experiment is not inexplicable. The null result for errors could be explained by a lack of reliability in the methodology as a whole, but its failures are underrepresented in the literature so it is impossible to know for certain. The competing errors were lacking due to shortened response deadlines, and the contextual effects of interruptions could be due to the amount of time each context requires to process initially, with more interruptions made in the lexical condition due to the speaker's haste, and combined with the monitor activity. And finally, while a lack of response to a target could be due to forgetting or selection difficulty, the context by outcome correlation when people did not respond could be due to spreading activation after an error creating either too few or too many possible options when the production process has to be restarted, resulting in more errors when the lexicality of the context and that of the intended error outcome match. A lack of precedence for investigating the interruptions across contexts as well as reporting the 'no response' results in any permutation of the SLIP task both indicate that further testing will be required to more fully explore the causes of these results.

7. References

Baars, B. J. (1980). On eliciting predictable speech errors in the laboratory. In V. A. Fromkin (Ed.), Errors in linguistic performance: Slips of the tongue, ear, pen, and hand (pp. 307–317). New York: Academic Press.

Baars, B. J., & Motley, M. T. (1974). Spoonerisms: Experimental elicitation of human speech errors. Journal Supplement Abstract Service, Fall 1974. Catalog of Selected Documents in Psychology, 3, 28–47.

Baars, B. J., Motley, M. T., & MacKay, D. G. (1975). Output editing for lexical status in artificially elicited slips of the tongue. *Journal of Verbal Learning and Verbal Behavior*, *14*, 382–391.

Blackmer, E. R., & Mitton, J. L. (1991). Theories of monitoring and the timing of repairs in spontaneous speech. *Cognition*, *39*, 173-194.

Del Viso, S., Igoa, J. M., & Garcı´a-Albea, J. E. (1991). Autonomy of phonological encoding: Evidence from slips of the tongue in Spanish. *Journal of Psycholinguistic Research*, *20*, 161–185.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*, 283–321.

Dell, G. S. (1990). Effects of frequency and vocabulary type on phonological speech errors. *Language and Cognitive Processes*, *5*, 313–349.

Dell, G. S., & Reich, P. A. (1981). Stages in sentence production: An analysis of speech error data. *Journal of Verbal Learning and Verbal Behavior*, *20*, 611–629.

Hartsuiker, R., Corley, M., & Martensen, H. (2005). The lexical bias effect is modulated by context, but the standard monitoring account doesn't fly: Related Beply to Baars, Motley, and MacKay (1975). *Journal of Memory and Language*, *52*, 58–70.

Hamm, S., Junglas, K., & Bredenkamp, J. (2004). The central executive as a prearticuliatory control device. *Zeitschrift für Psychologie*, *212*, 66–75.

Haupt, E. (2001). The First Memory Drum. *The American Journal of Psychology*, *114*(*4*), 601-22.

Humphreys, K. R. (2002). Lexical bias in speech errors. Unpublished Doctoral dissertation, University of Illinois at Urbana-Champaign.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1–75.

Nooteboom, S. (2005). Listening to one-self: Monitoring speech production. R.J. Hartsuiker, Y. Bastiaanse, A. Postma, and F.N.K. Wijnen (eds.) "Phonological encoding and monitoring in normal and pathological speech", pp. 167-186.

Nooteboom, S., & Quene´, H., (in press). Self-monitoring and feedback: A new attempt to find the main cause of lexical bias in phonological speech errors. *Journal of Memory and Language* (2007).

Oomen, C. C. E., & Postma, A. (2001). Effects of time pressure on mechanisms of speech production and self-monitoring. *Journal of Psycholinguistic Research*, *30*(*2*), 163-184.

Postma, A. (2000). Detection of errors during speech production: a review of speech monitoring models. *Cognition*, *77*, 97-131.

Van Hest, G. W. C. M. (1996). Self-repair in L1 and L2 production. Tilburg: Tilburg University Press.

Acknowledgements

I would like to thank Martin Corley and Suzy Moat for their invaluable input throughout the creation and execution of this dissertation, Cara Featherstone for her skills with a synthesizer, Judith Köhne, Nien-Chien Lee for their comments, and Paul Riddle for editing and, most importantly, his unending support.